MACHINE LEARNING NANODEGREE PROGRAMME

Capstone Proposal

David K. Njuguna
(dakn2005@gmail.com)
August 14th, 2018

Abstract

The proposal is to create a model for classifying households eligible for cash transfer using provided household features. This is done by grouping household in either *group 1* for households which are marked as very vulnerable and are eligible (in this case) for bi-monthly payments, or *group 2* for households marked to be paid in emergency cases only and are not as vulnerable.

To achieve this, we'll be comparing the actual grouping classification performed in the dataset with the grouping classification performed by the model.

Domain Background

Cash transfer programmes are designed to help alleviate the individuals from poverty[6]. These programmes are funded mostly by governments[5] under Social Protection Services[4], but are not limited to being government-funded.

Non-governmental organizations can also operate some cash transfer programmes, depending with the social protection laws implemented in individual countries.

As an information systems practitioner, my major motivation of implementing this project is providing a model enabling NGOs with a quick way of determining eligibility of a household using selected features. This is because the process of determining eligibility can be very time consuming and costly, the after effect of this being communities in need having an unnecessarily lengthy wait time before aide arrives.

I'll be using supervised learning algorithms. Some of the algorithms I'll be testing to produce the best model are decisiontreeclassifier, SVM and XGBoost

Problem statement

This is a classification problem, where a household is to be classified into two groups. These groups determine the household payment cycles, depending on the severity of the household (from poor to better off) and other features which will be investigated.

Datasets and inputs

The anonymized dataset can be accessed via https://github.com/dakn2005/MLNDProject_CT/blob/master/mlnd-ds-2.csv.

To acquire/request for new data, you have to send a request form to the programme using http://hsnp.or.ke/images/mis/hsnp_data_request_form.pdf

Anonymized dataset has 100,000 households, with the following fields:

Categorical variables: Gender, Attended_School, Work_last_7days, Chronic illness, Disabled, Wealthgroup (decided using community based ranking), Resident_Provider, HHGroup (household group), Toilet (source), Drinking_water (source)

Discrete variables: Children_under_15, Kids_under_15_in_settlement, wives_in_settlement, spouses_outside_hh, Land_ha (hectarage)

Continuous variables: age, years of chronic illness

I will segregate the data into training set (80%) and testing set (20%)

The target label is **HHGroup**, with around 27% households classified in group, while the rest 72% are classified in group 2. The dataset is therefore imbalanced.

Also the dataset has a large amount of missing data marked as 'SKIP'. All missing data will be removed from the dataset in the data preparation stage.

Solution statement

Using various machine learning algorithms to perform these classifications using the input features. Use of classification algorithms i.e. decisiontreeclassifier, svm or xgboost to best classify a new household given the input features.

Benchmark model

Will implement a *naive predictor* as the base model, with the assumptions that no negative values are predicted. Will check out the accuracy, f-score, recall and precision values. The resultant final model should give better scores especially on accuracy in comparison with this model.

Evaluation Metrics

Will compare model accuracy and f-beta score investigating both precision and recall , with emphasis on recall. While the accuracy metric is important, will be checking for precision and recall due to the dataset being imbalanced. Of particular interest is checking the sensitivity of the resultant model (how many relevant households are selected).

tp = true positive, tn = true negative, fn = false negative, fp = false positive, N = dataset size, B = beta

Recall = $\frac{tp}{tp + fn}$

Precision = $\frac{tp}{tp + fp}$

Accuracy = $\frac{tp + tn}{N}$

$F_B = (1 + B\text{^}2).\frac{precision.recall}{(B\text{^}2.precision) + recall}$

Project design

First, we'll perform some data processing steps. These include: removing missing data, normalization of numeric figures, checking if there are features that will require tranformation, encoding categorical variables (using one hot encoding scheme) to convert non-numeric values to numeric values, splitting the dataset to training and testing sets

Secondly, we will set the benchmark model using the naive predictor as mentioned.

Thirdly, we'll test out the different classification algorithm checking for time taken and the accuracy metrics of individual algorithms on the testing and training sets. The tests will be performed on different sizes of data starting with 1%, then 10% and finally 100% of data. I have chosen three common and most prevalent supervised algorithms to be tested.

Fourthly, select the most performant model comparing time taken for execution versus scores produced. Ideally the best model should use the least amount of time whilst produce good(acceptable) scores

Fifth, investigate on feature importance, and impact of the model selected with fewer features, selecting the features marked as important only

Tools used: python 3.x, jupyter notebook, sklearn, pandas, numpy

References

1. Hunger Safety Net Programme, Cash Transfer Program

http://hsnp.or.ke/

2. Supervised Learning

https://en.wikipedia.org/wiki/Supervised_learning

3. Precision and recall

https://en.wikipedia.org/wiki/Precision_and_recall

4. Why we need social protection

http://www.developmentpathways.co.uk/publications/why-we-need-social-protection/

5. HSNP2 Impact Evaluation Final Report

http://hsnp.or.ke/index.php/our-work/downloads

6. Cash transfers: what does the evidence say?

https://www.odi.org/sites/odi.org.uk/files/resource-documents/10749.pdf